Measurement and Fuzzy Scales[1]

Misurazione e scale sfocate

Michele Lalla
Università di Modena e Reggio Emilia, Dipartimento di Economia Politica, Via J. Berengario 51, 41100 Modena, Italy, e-mail: lalla.michele@unimore.it

Gisella Facchinetti
Università di Modena e Reggio Emilia, Dipartimento di Economia Politica, Via J. Berengario 51, 41100 Modena, Italy, e-mail: facchinetti.gisella@unimore.it

Riassunto: la misura dei concetti presenta diverse difficoltà e gli strumenti utilizzati, per rilevare caratteri qualitativi ordinati, non sono sempre soddisfacenti. Tra le scale ordinali si descrive quella di Likert e tra le alternative si delinea la possibilità di applicare la metodologia degli insiemi sfocati per ottenere valori individuali che siano prossimi alla realtà e coerenti con il processo di misura. Si illustrano, quindi, i risultati di una indagine condotta per accertare la differenza tra due scale di valutazione della didattica: una proposta dal Comitato nazionale per la valutazione del sistema universitario e l'altra corrispondente ai voti tradizionali attribuiti nella valutazione dell'apprendimento nelle scuole precedenti l'università (scala di voto). Gli esiti mostrano che questa sembra piú coerente con i punteggi attribuiti alle modalità.

Key words: ordinal measurement, Likert scale, neutral position, fuzzy set theory

1. Introduction

A number of different types of rating scales or scale formats are available to measure the intensity of concepts or attitudes (*e.g.* semantic differential, Stapel scale, Likert scale, Thurstone scale, and the direct rating scale). They generate numbers that represent a rough ordinal level of the attribute at the most, while data processes involve indices and parameters implying that the resulting scores are real numbers. In fact, there are many approaches to measure the direction and strength of an attitude.

The Italian Committee for University System Evaluation proposed a course-evaluation questionnaire (Chiandotto and Gola, 2000) containing items with a four-point Likert scale: ①*Definitely not*, ②*No, rather than yes*, ③*Yes, rather than no*, ④*Definitely yes* (referred to as the CUSE scale). They also suggested using means and variances to analyse data, translating each category into the *decimal scale* values ranging from 0 to 10: the complete set was $\{2,5,7,10\}$. On the one hand, the CUSE scale does not have a middle category, it may violate the linearity assumption, and the numbers are assigned to alternatives almost arbitrarily. On the other hand, mean and variance cannot validly be used with it. An alternative set of categories is: ①*Very insufficient*, ②*Insufficient*, ③*Sufficient*, ④*Good*, ⑤*Very good* (referred to as the mark

[1] Financial supported was provided by the Italian MIUR for the project, «Industrial Districts as Complex Systems», approved in 2002 (prot. 2002133972) and directed by David Avra Lane.

scale) because it is fairly similar to the grading system used at previous school levels where many decimal scale ratings are expressed through these syntagms. Therefore, they are recognised by all students, although other terms are also used in which ambiguities and differences still persist and may vary from school to school, as is the case with the "mediocre" and "fair" (*discreto*) ratings. Each item score could be translated into a decimal scale score by multiplying the numerical label of the category by two, with the complete set being $\{2,4,6,8,10\}$. Our remarks on the CUSE scale remain valid for the mark scale, although the latter has peculiar characteristics that should make it work better than the former.

The first aim of this paper is to describe the general characteristics of these devices as related to teaching evaluations. The description of the questionnaires used to survey the opinions of students, the difficulties involved in the two Likert-type scales, and some empirical results are illustrated in Section 2. Secondly, a Fuzzy System (FS), based on the Fuzzy Set Theory (FST) introduced by Zadeh (1965), is presented as an alternative strategy to analyse qualitative ordinal data and is then compared with the traditional one. Some aspects of FST of use to the social sciences, an FS built up to generate the numerical evaluation for some conceptual domains of teaching activity and some possible developments in scaling methods and data analysis are briefly described in Section 3. Comments and possible further developments are reported in Section 4.

2. Likert-type scales and the evaluation of teaching activity

The Likert scale is used to measure attitudes and opinions through *L* statements, supposed to be semantically connected with the surveyed object or concept, as in other scales such as the semantic differential, feeling thermometers, Thurstone, and Guttman. Each statement (item) has favourable or unfavourable contents to a concept or attitude with a varying degree of intensity and the *L* statements should form two sets, both having almost the same cardinality. They should be monotone, *i.e.*, unidirectionally maintained with respect to the measured object to yield an increasing score as the attitude of a subject increases. Each item offers respondents these options: *strongly agree*, *agree*, *uncertain*, *disagree*, *strongly disagree*. Each subject expresses his/her agreement with the contents of the statements by choosing one of them. The respective scores are $\{5,4,3,2,1\}$ and $\{1,2,3,4,5\}$ for favourable and unfavourable statements regarding the concept. The final score, y_i, for the *i*-th subject is $y_i = \sum_{l=1}^{L} y_{il}$.

The format, namely the number of categories, is arbitrary. In fact, the alternatives were originally seven. It was decreased to five to simplify the set of options, to reduce the variability of the zero-point and the unit of measurement of individuals. However, the (optimal) number of alternatives is a function of measurement conditions and the specific meaning of the question (Mattell, Jacoby, 1971; Cox, 1980; Wildt, Mazis, 1978). This procedure facilitates administration (of the test), registration, coding, and processing. It therefore enables the researcher to collect a large amount of data in much less time and with the same set of possible alternatives. It assumes unidimensionality of attitudes, location of attitudes on a *continuum*, and equidistance between alternatives. However, the construction strategy does not guarantee the measure of only one property, nor the equality of the perceived positions for each alternative and for all subjects (Duncan, Stenbeck, 1987; Brody, Dietz, 1997).

The Likert scale poses, in fact, many problems. The presence/absence of the neutral point —denoted by syntagms such as «same as now», «right amount», «I don't mind»— is a debated issue. It is often eliminated to press respondents to choose a sharp alternative instead of allowing refuge in a middle position, assuming that (1) it attracts people who have no opinion or prefer a noncommittal position rather than saying «I don't know», (2) respondents tend toward one or the other polar alternative, (3) people who really are neutral, randomly choose one of the two nearest alternatives (Schuman, Presser, 1996). A scale without a middle position has no position equal to "zero" and the alternative options are no longer equidistant, but its inclusion does not solve the difficulties of using option values as real numbers. Therefore, the responses collected could misrepresent or bias the actual attitudes of respondents (Guy, Norvell, 1977; Ryan, 1980). In fact, the percentage for the middle position decreases when the number of alternatives increases (Mattell, Jacoby, 1972), although respondents are affected by the nature of the questions. The Likert-type scale also presents many other difficulties. First, the closed-question format obliges respondents to choose only from among the available options, which may not match their real opinions or attitudes. This leads to an increase in missing data and possible drifting toward the social acceptability of the answers varying between individuals, over space, and time (Orvik, 1972). Second, the assumptions of continuity and equidistance of the alternatives are arbitrary, as they vary among individuals without plausible regularity and there is an end effect. Third, the linearity assumption could be violated, generating curvilinearity, when two persons, having opposite attitudes, give the same answers for opposite reasons. Therefore, truly unfavourable people are joined with those who are more favourable towards the concept than the intensity expressed by the statement (Coombs, 1953). Fourth, the answers could be affected by (acquiescent) response set or proximity, especially in a battery of questions, which is also referred to as *yea-saying* or *nay-saying* for dichotomous alternatives, or the *donkey vote effect* for the choice of the first alternative. These are form-related errors related to the psychological orientation of respondents, such as proximity, leniency, central tendency, i.e., a propensity to rate something/someone too high (or too low) or reluctancy to choose extreme values (Albaum, 1997). Fifth, the position on the page and the order of the alternatives could generate the *primacy effect*, i.e., respondents tend to select items placed on the left side of the page or the alternatives on the left side of the scale (Chan, 1991). In teaching evaluations, this effect shifts the mean rating to the lower end when the positive end of the scale is on the right (Albanese *et al*., 1997). Sixth, the statements could generate a reaction to the "object" involved in the expression, as in the case of an answer affected by the context and not by the meaning. Therefore, a respondent will always answer a question favourably, when he/she has a favourable attitude towards the object involved in a statement, even if he/she is not in favour of the actual content (Cacciola, Marradi, 1988).

There are many alternatives to these scales. For example, in feeling thermometers, it is possible to increase the number of options or to fix a segment where the respondent marks his/her position. In the latter, the distance of the mark from the origin is the measure of intensity and the outcome is a real number, but the evaluations of respondents are more approximated than the values obtained. In fact, Hofacker (1984) observed that respondents only used about eight answer categories, although the scale adopted allowed them to express very fine-grained distinctions. Marradi (1992, 1998) argues that a better approximation to cardinality is obtained using *self-anchored* scales, which reduce the semantic independence of the intermediate options, but respondents

tend to use only multiples of 10, as the range is 0-100. Therefore, a more finely-grained structure of the scale does not ensure better precision.

Three split-ballot experiments were carried out to obtain student ratings for each labelled category and evaluate the performance of the mark and/or CUSE scales. Eight key items were selected from the questionnaire proposed by Chiandotto and Gola: Adequacy of the Lecture Room (1.ALR), Adequacy of the Work Load requested (2.AWL), Correspondence between Actual and Planned lectures (3.CAP), Adequacy of Teaching Materials (4.ATM), Clarity of the Teacher's Presentations (5.CTP), Teacher Availability during Office hours (6.TAO), Usefulness of Teaching-Support Activities (7.UTS), Level of Interest in the Subject matter (8.LIS). Three different questionnaires were constructed: Q1, Q2, and Q3. Each one contained the eight items, but the answers were graded differently: Q1 used the mark scale, Q2 the CUSE scale, and Q3 used both. The three questionnaires were administered to students attending their first, second, or third year, respectively. The respondents had to specify their ratings for each categorical alternative for each item, using a decimal scale ranging from 0 to 10. Details can be found in Lalla *et al*. (2004).

The survey sampling sizes are reported in Table 1 and the figures reveal a better performance of the mark scale, which showed a low error rate. The types of errors, which are not reported here (Lalla *et al*., 2001), suggested that the CUSE scale could be misleading for respondents. The percentages of incomplete evaluations of the options were higher for Q2 than for Q1. The percentages of inconsistency (a sequence of values attributed to the five options which is extremely non monotonic) were higher for Q2 than for Q1 again, while the differences were reduced in Q3, suggesting that the mark scale acted as a guide in attributing values to the labelled options on the CUSE scale.

The main results, as reported in Lalla *et al*. (2004), are the following. The first-year students were in the three Mathematics classes and evaluated the mark scale categories (Q1) using the decimal scale. The results, including those of Q3, are shown in Figure 1. The means of the decimal scores proved to be higher than the value assigned to the categories $\{2,4,6,8,10\}$, for those categories below "sufficient". They were lower than the value assigned to the categories, for those categories above "sufficient". There was also heteroscedasticity: the standard deviations decreased for increasing levels of the scale, up to sufficiency, and slowly increased for increasing levels above sufficiency. The standard deviations in the first level (or option) were generally higher than others.

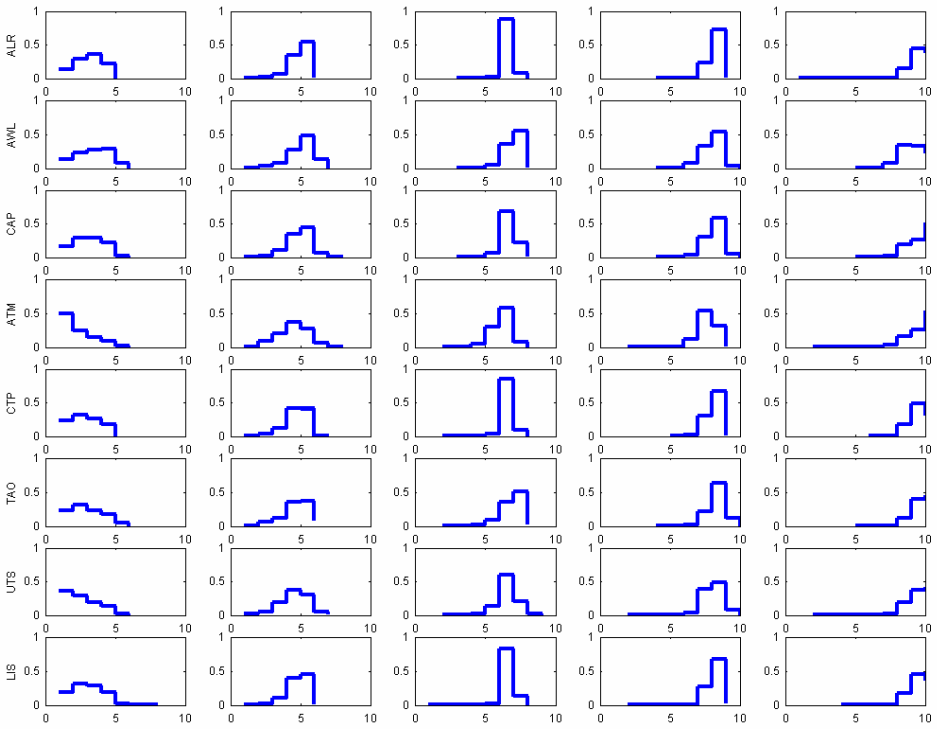
Table 1: *Number of cases and percentages by type of questionnaire and consistency of answers with respect to the evaluation of the categories*

Questionnaire	**Consistent**	**%**	**Inconsistent**	**%**	**Total**
Q1 Total	245	70.6	102	29.4	347
Q2 Total	110	45.6	131	54.4	241
Q3 Total	146	72.0	57	28.0	203
(Q1+Q2+Q3) Total	**501**	**63.3**	**290**	**36.7**	**791**

The second-year students were in two separate Political Economy 2 classes and two Mathematics for Financial Market classes. They evaluated the CUSE scale categories (Q2) using the decimal scale again. The results, including those of Q3, are shown in Figure 2. The means of the decimal scores were markedly lower than the values attributed to the labels of the categories. Differences proved to be less accentuated only for «*Definitely not*» and «*Yes, rather than no*». There was heteroscedasticity without a

specific pattern. The standard deviations for the CUSE scale were higher than those for the mark scale. These findings could denote a better performance of the mark scale and more uncertainty among students in attributing values to CUSE scale options.

Figure 1: *Relative frequency histogram (stair-step graph) of decimal scores attributed to mark scale options by students (data gathered through questionnaires Q1 and Q3)*

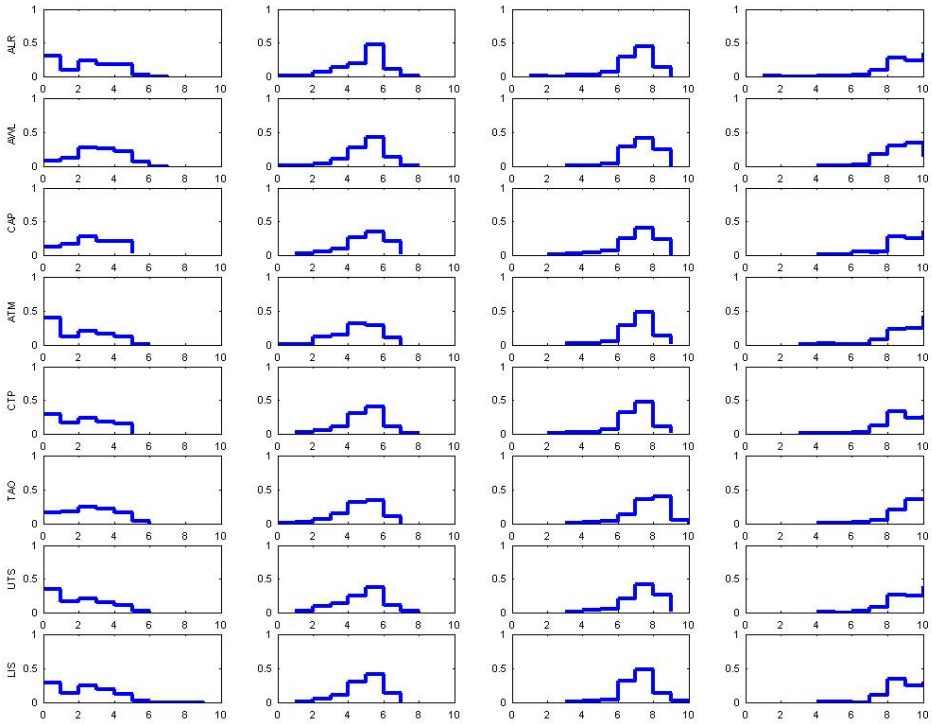


The third-year students (evaluating both the mark and CUSE scale categories in Q3, using the decimal scales) were attending the three separate Public Economics classes and two Banking and Finance classes. The observed mark scale means proved to be statistically equal to the corresponding means of values collected with Q1, whereas the observed CUSE scale means proved to differ statistically from the corresponding means of values collected with Q2. Furthermore, the evaluations for the mark scale categories neared those of the CUSE scales, suggesting that the mark scale gave respondents its cue in evaluating the intensity of the CUSE scale options. The mark scale means were slightly higher than those for the CUSE scale, while heteroscedasticity persisted in the data. Furthermore, the Kolmogorov-Smirnov tests showed that the distributions of the attributed values for the mark scale options surveyed with Q1 were statistically equal to those surveyed with Q3, whereas the distributions of the attributed values for the CUSE scale options collected with Q2 were completely different from those collected with Q3. Therefore the most suitable and reliable scale is the one already well known to the target population, the mark scale which is used in secondary schools.

The evaluations of the scale options were analysed through a model for repeated measurements, to ascertain the differences between the mark and CUSE scales and the influence of concomitant variables, such as gender and teacher (Lalla *et al.*, 2004). Referring to the CUSE scale, the Q2 and Q3 profiles were not parallel even in the case of

items having the same type of labelled options. Therefore, this empirical evidence could suggest that the CUSE scale yields unstable results and that the stimuli affect subjects less uniformly than mark scale stimuli, as the heterogeneity of the variances was greater in the former than in the latter.

Figure 2: *Relative frequency histogram (stair-step graph) of decimal scores attributed to CUSE scale options by students (data gathered through questionnaires Q2 and Q3)*



The effects of other factors were analysed for the mark and CUSE scales, separately. Gender showed a heterogeneity of variances and interaction with the level of the scale, while no effects were revealed in relation to the type of education or interaction with gender. It was ascertained that the values for the categories of the scales depended on item wording, the terms used to define the five or four categories (labels) or the item itself, the nature of the question, the type of course, the teacher, and the intensity level of the option. These patterns were similar for both scales, but the differences were higher for the CUSE scale and verified at any level (option) *k*.

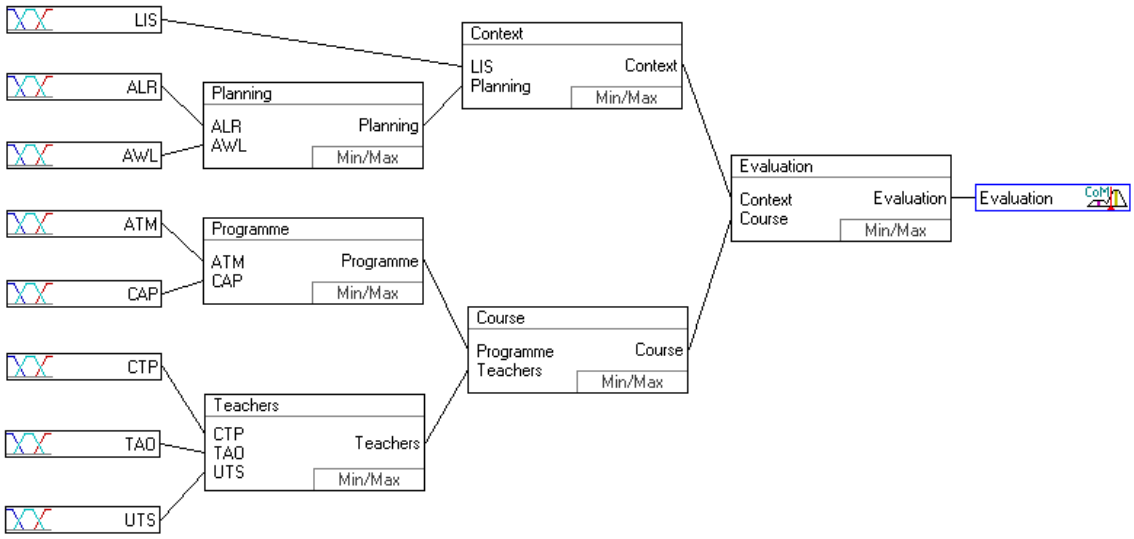
3. The fuzzy set theory

The concept involved in FST is *fuzziness*, which refers to the ambiguity, impreciseness, indistinctness, inexactitude, vagueness, and similar attributes that characterise current propositions and human thought. The belonging to a category (set) is described by a membership function, *i.e.*, instead of an object either belonging to, or not belonging to it, as in classical set theory, the object may be *more* or *less* its member. In fact, the grade of membership function is a (subjective) measure of affiliation of an element with the

set. The opposite of fuzziness is *crispness*, referring to clearness, exactitude, distinctness, and preciseness set membership, *i.e.*, the numerical value quantifying the strength of belonging to a set. FST is described in many textbooks (Zimmermann, 1996; Dubois, Prade, 2000). Therefore, how an FS works and its potentialities are described by setting it up in a practical application. An FS is generally designed through the following sequential hierarchical steps (Kasabov, 1996).

The *identification of the problem*, (*i*), requires an in-depth analysis of the problem to clarify the concepts, to develop operational definitions, and to outline the set of relationships among the variables. It culminates in a sort of forward amalgamating tree, as shown in Figure 3 for the eight items relating to teaching activity. Several fuzzy modules are linked together and the variables are introduced at different levels of importance, while in the arithmetic mean they are weighted equally. Each single aggregation produces intermediate variables that have a particular meaning. For instance, the aggregation of ALR (Adequacy of the Lecture Room) and AWL (Adequacy of the Work Load requested) generates the new variable, «Planning», which summarizes all the information about the Faculty organisation (Lalla *et al*., 2001).

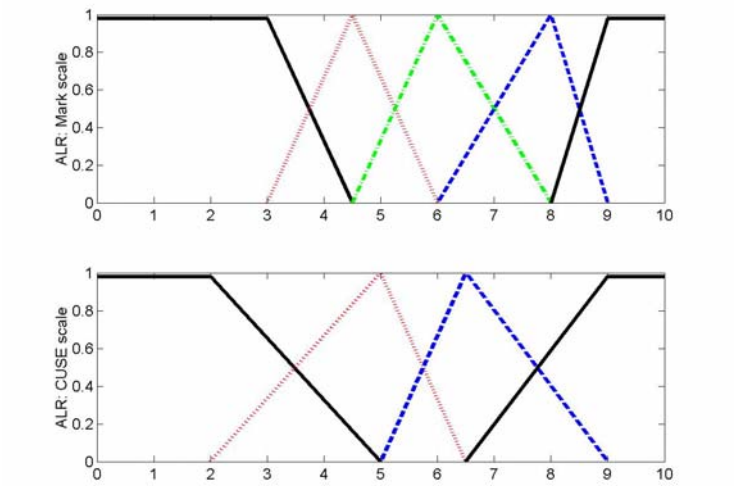
Figure 3: *A hypothetical fuzzy system for teacher and course evaluation*



The *fuzzification of input variables*, (*ii*), concerns the definition of the membership function for each category of a variable and for all variables. There are several approaches: formalist, comparative, scaling, pollster (Smithson, 1988, p. 5), *knowledge-based*, and *machine-learning* methods. The pollster method is more suitable for teaching evaluation as it requires a survey of a target population sample to obtain a relative frequency distribution of the scores collected for each option of a variable and for all variables. The shapes of the empirical distributions will suggest the membership functions. In fact, here the fuzzification of input variables was carried out by examining Figures 1 (or Figure 2) to identify the form, peak, and amplitude of the corresponding fuzzy number. Almost all the terms were represented by piece-linear functions. For example, Figure 4 illustrates the fuzzification of ALR for mark and CUSE scale, where the triangular or trapezoidal lines are traced through the points determined by the medians of scores attributed by students to the ALR options for both the mark and CUSE scales. This procedure was repeated for each one of the seven remaining items.

The *construction of rule-blocks*, (*iii*), can also be obtained in several ways. Here, the rule-blocks were set up through the experts' opinions. The formulation of a rule connects the antecedent with the consequent, through an "IF-THEN" statement. In each block of the design outlined in step (*i*), the "IF-THEN" connects every combination of the input option variables with an output variable option. For example, «IF ATM is ①*Very insufficient* and CAP is ①*Insufficient* THEN the output of the "Programme" block (see Figure 3) is ①*Insufficient*». Intermediate blocks had an output variable with five possible options, while the final block had seven. This is an onerous step as each aggregation module requires a rule for each combination of the possible outcomes for all the input variables in the module. Therefore, just two variables with five alternatives each lead to 125 combinations, *i.e.*, 125 rule-blocks connecting the input to the (intermediate) output, and three variables lead to 625 (five to the third) combinations. Setting up the rules is a delicate matter because the choices adopted are determinant, but also extremely subjective. In fact, to state a sensible rule, the operator should be aware of the problem and be an expert: the more capable the operator, the better the FS will work and reach a "good solution". Therefore, it cannot be used to find the optimal solution, typical of classic control theory. The optimal in the real world represents a difficult problem, subject to doubtful interpretations, precisely according to the basic ideas of FST (Bojadziev, Bojadziev, 1997).

Figure 4: *Hypothetical membership functions for Adequacy of the Lecture Room*



The *aggregation of rule-blocks*, (*iv*), involves the handling strategy of the input fuzzy sets to produce a fuzzy region representing the intermediate or final output (target) variable. This is the working core of the FS, which incorporates the weights activating each single rule. Again, there are many methods to aggregate variables both in the precondition (IF) and in the conclusion (THEN). In the precondition part, there are several operators (*t-norms*) that perform intersections between fuzzy sets, such as the MIN, MAX, their convex combination ($\lambda \cdot \mathrm{MIN} + (1-\lambda) \cdot \mathrm{MAX}$ with $\lambda \in [0,1]$), and γ-operator. The latter represents a degree of positive compensation, defined by:

$$\mu = \left(\prod_{l=1}^{L} \mu_l \right)^{(1-\gamma)} \left[1 - \prod_{l=1}^{L} (1-\mu_l) \right]^{\gamma}. \qquad (1)$$

The choice of the γ parameter requires an adaptation to real data and the empirical range varies between 0.25 and 0.4 (von Altrock, 1997). The γ-operator satisfies many

conditions compensating the excess of satisfaction for one operator and the exiguity of satisfaction for the other one. In the conclusion part, the popular operators are the MAX that keeps the strongest rule as "winner" and the BSUM (boundary sum) that keeps the whole firing degree, and the final result is the sum of the different levels of activation (not over 1).

The *defuzzification of output*, (*v*), pertains to the transformation of the output fuzzy sets into a crisp value. The selection of a proper method requires an understanding of the linguistic meaning and the workings of the different available procedures that are, for example, the Centre of Area (CoA), the Centre of Gravity (CoG), the Mean of Maximum (MoM), and the Centre of Maximum (CoM). Processing and defuzzification of data from the evaluations of teaching activity were carried out for each respondent using the support of "fuzzyTECH" by INFORM. The crisp value was obtained by the CoM method, which determines the most typical value for each term, being the best compromise for the fuzzy inference resulting from their weighted average. The weights were the levels of activation of the geometric figure resulting from the union of the output fuzzy sets.

A *sensitivity analysis*, (*vi*), is necessary to identify the more suitable membership functions, rules, operators, and techniques (pertaining to each previous step) to achieve the best performance of the FS in the description of the phenomena under examination.

The survey using the questionnaires Q1, Q2, and Q3 was aimed mainly at obtaining the membership functions and ascertaining the differences between the CUSE and mark scales. To verify the performance of the previous FS, it is necessary to have the actual courses evaluations. The University of Modena and Reggio Emilia used a questionnaire with 24 items (inclusive of the 8 items listed above) concerning teaching activity that are not reported here, for the sake of brevity. Each item used a mark scale, in contrast with the indications of the Italian Committee for University System Evaluation. At this stage, only items selected for evaluations and statistical units referring to designated courses were extracted from the archive containing data gathered in December 2000.

A first rough comparison between the ordinal and fuzzy ratings was made by dichotomising the scores: «insufficient» less than six, and «sufficient» otherwise (Lalla *et al*. 2004). The ratings agree when both mean scores yielded the same judgment, *i.e.*, the answers corresponding to sufficient on the decimal scale were sufficient also with the fuzzy score, and likewise for insufficient scores. There was agreement in 80% of cases: 70% were sufficient, while only 10% were insufficient. There was disagreement in the remaining 20% and, surprisingly, the means of the decimal scale all proved to be sufficient and the means of fuzzy scores all proved to be insufficient, which indicated a systematic underestimation of the numerical values in the fuzzy evaluation, although the fuzzy scores were higher than the traditional ones in 45% of the cases. The correlation coefficients for the ordinal and fuzzy ratings were stable within classes (about 0.89), but the fuzzy scores over-estimated high means and underestimated low means. In fact, they tended to expand the values over the entire range of the decimal scale, lowering low scores and raising high scores, which generated a weak S-shaped nonlinearity. This tendency depended on many factors, as outlined in step (*vi*), such as the aggregator operators selected in the intermediate steps, the membership functions designed for the output options, and the defuzzification method adopted.

The ordinal rating was also obtained using the mean derived using the pollster method for each corresponding option of the mark scale, instead of using $\{2,4,6,8,10\}$, but the results only showed a generalised decrease of means within classes, although the

differences rarely exceeded one half point in the worst cases (Lalla *et al*., 2004).

An FS requires the aggregation of at least two variables to be applied, but the analysis of students' judgments can require the examination of a single item because a teacher might also want a specific item evaluation, *i.e.*, an item-by-item analysis, which is not feasible through a standard FS. Two main issues need to be solved within this framework. The first involves a sort of aggregation or a specific strategy to allow the FS to work. The second concerns the spreading of the distribution of scores gathered through the pollster method on the unevaluated items of a battery, as it is not possible to submit the sampling units to the same format repeatedly.

The analysis of single items requires a corresponding crisp value for each option. This entails the application of an FS to one input variable and the fuzzification of the single item, which will define a numerical output for each option. As two membership functions (of the two corresponding options of an item) are overlapped in some regions, the fuzzy representation of each crisp input value, attributable to a response to that item, always produces a new figure built by the union of the two linguistic attributes that the crisp value fires, with cut-off at the level of activation of the single attributes. Through defuzzification, (v), of that figure (one for each option), one-to-one correspondence is obtained between the defuzzified values and the option values. The procedure should be applied to each item involved in the FS.

Further developments concern alternative strategies to obtain the global response of an FS. For example, the FS now works by aggregating the responses of the same individuals and the final response is given by an arithmetic mean of the single defuzzified output. Therefore, two different aggregations of the (final) outputs of different individuals are feasible, referred to as the "*a-union*" and "*a-sum*". The *a-union* is the value obtained by the defuzzification of the union of the n (total number of cases) output fuzzy sets obtained. As is known, every *t-conorm* is a way to evaluate the union of fuzzy sets. The response (variable) depends on two parameters: the defuzzification method and the *t-conorm* that will be chosen. The *a-sum* is the usual definition of sum in FST. Furthermore, the same problem arises in the evaluation of a single item. For its final evaluation, it is enough to obtain the frequency distribution per item, *i.e.*, the number of respondents for every option (or single linguistic attribute) and to determine an average. In the ordinary case, the single value label is multiplied by its relative frequency and then the products are summed. In the fuzzy case, one proposal is to multiply every fuzzy number (describing every linguistic attribute) by the relative frequency of the option (linguistic answer) and thus obtain five fuzzy sets. At this stage, it is possible to work, as above, by an a-union or an a-sum aggregation. Therefore, the single-item final response is given by the defuzzification of the a-union or a-sum of the fuzzy set obtained. The a-sum is the «exact» translation of the ordinary weighted average and the response (variable) depends only on the defuzzification method used.

4. Conclusions

The FS is a flexible method to obtain a numerical value from ordinal variables, such as course evaluation data, and presents no problems regarding some difficulties typical of many scales such as the issue of a middle position, which vanishes in the FS operating procedure. FS design could easily incorporate the knowledge of an expert in the fuzzification of input and in the building-up of control block-rules, as they allow for the

best adaptability. These are strengths, but also weaknesses as they involve extremely subjective, ambiguous, and «private» decisions, when scientific procedures should be objective, unambiguous, and public. Actually, the architecture of the FS could be public, but, for example, the number of rules grows exponentially and their control is arduous. Furthermore, the aggregation of variables proceeds through a tree, varying nonlinearly their impact on the output, as their influence depends on the levels of the knot where they enter. The responses are a sort of weighted average with «unknowable» weights. However, in spite of these difficulties, the FS could be a valid and reliable tool to represent situations described by qualitative ordinal variables comparable with others.

The FS does not perform single-item analysis well, as it works on an aggregation of a small number of variables (as few as two or three). To extend the applicability of an FS to teaching evaluations, an item-by-item analysis may be feasible, but the proposed procedure is still too simple to quantify the amount of vagueness in the intensity expressed by the option label and very similar to a common ordinal scale.

Many popular scales, indexes, and measures may be conceived as fuzzy sets denoting graded concepts. Therefore, FST could provide an alternative approach to measurement for the social sciences and, although it showed comparable results, it is in need of improvement and extension. The combination of computational intelligence with FS could represent a useful route for the analysis of individual behaviour or judgments. Human action is characterised by great complexity as it involves a lot of variables (some of which are latent, omitted, and not always exactly measurable), spatial, individual and/or group heterogeneity and contamination between the actors. Soft computing techniques could be an interesting strategy to model social actions because they allow for a representation of reality in its wholeness, without introducing restrictions or reductions, even if they still require further refinement and adjustments to compete with traditional methods.

References

Albanese M., Prucha C., Carnet J. H., Gjerde C. L. (1977). The Effect of Right or Left Placement of the Positive Response on Likert-type Scales Used by Medical Students for Rating Instruction, *Academic Medicine*, 72, 627-630.

Albaum G. (1997). The Likert scale revisited: an alternate version, *Journal of the Market Research Society*, 39(2), 331-348.

Bojadziev G., Bojadziev M. (1997). *Fuzzy logic for business, finance and management*, World Scientific, Singapore.

Brody C. J., Dietz J. (1997). On the dimensionality of two-question format Likert attitude scales, *Social Science Research*, 26 (2), 197-204.

Cacciola S., Marradi A. (1988). Contributo al dibattito sulle scale Likert basato sull'analisi di interviste registrate, in Marradi A. (ed.), *Costruire il dato. Sulle tecniche di raccolta delle informazioni nelle scienze sociali*, Franco Angeli, Milano, 63-102.

Chan J. C. (1991). Response-order Effects in Likert-type Scales, *Educational Psychology Measurements*, 51, 531-540.

Chiandotto B., Gola M. M. (2000). Questionario di base da utilizzare per l'attuazione di un programma per la valutazione della didattica da parte degli studenti, *Rapporto finale del gruppo di Ricerca* (RdR 1/00): MURST, Osservatorio per la valutazione

del sistema universitario (*http://www.cnvsu.it*).
Coombs C. H. (1953). Theory and Method of Social Measurement, in Festiger L., Katz D. (eds.), *Research Methods in the Behavioral Sciences*, Dryden, New York, 471-535.
Cox E. P. (1980). The Optimal Number of Response Alternatives for a Scale: A Review, *Journal of Marketing Research*, 17, 407-422.
Dubois D., Prade H. (eds.) (2000). *Fundamentals of Fuzzy Sets*, Kluwer Academic, Boston.
Duncan O. D., Stenbeck M. (1987). Are Likert Scales Unidimensional?, *Social Science Research*, 16, 245-259.
Guy R. F., Norvell M. (1977). The Neutral Point on a Likert Scale, *The Journal of Psychology*, 95, 199-204.
Hofacker C. F. (1984). Categorical Judgment Scaling with Ordinal Assumptions, *Multivariate Behavioral Research*, 19(1), 91-106.
Kasabov N. K. (1996). *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering*, MIT Press, Cambridge, MA.
Lalla, M., Facchinetti, G. & Mastroleo, G. (2001). A fuzzy expert system for evaluating teaching efficiency, in Reinhard N. (2003), *Modeling and Control of Economic Systems* 2001, Elsevier (Pergamon Press), Oxford, 395-400.
Lalla M., Facchinetti G., Mastroleo G. (2004). Ordinal Scales and Fuzzy Set Systems to Measure Agreement: An Application to the Evaluation of Teaching Activity, *Quality & Quantity*, forthcoming.
Marradi A. (1992). *L'analisi monovariata*. Franco Angeli, Milano.
Marradi A. (1998). Termometri con vincolo di ordinabilità: il «gioco della torre» consente di aggirare la tendenza alla desiderabilità sociale?, *Sociologia e ricerca sociale*, 57, 49-59.
Matell M. S., Jacoby J. (1971). Is There an Optimal Number of Alternatives for Likert Scale Items? Study 1: Reliability and Validity, *Educational and Psychological Measurement*, 31, 657-674.
Matell M. S., Jacoby J. (1972). Is There an Optimal Number of Alternatives for Likert Scale Items? Effects of Testing Time and Scale Properties, *Journal of Applied Psychology*, 56 (6), 506-509.
Orvik J. M. (1972). Social Desirability for Individual, his Group, and Society, *Multivariate Behavioral Research*, 7, 3-32.
Ryan, M. (1980). The Likert Scale's Midpoint in Communications Research, *Journalism Quarterly*, 57 (2), 305-313.
Schuman H., Presser S. (1996). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*, Sage, Thousand Oaks, CA.
Smithson M. J. (1988). Fuzzy set theory and the social sciences: the scope for applications, *Fuzzy Sets and Systems*, 26, 1-21.
von Altrock C. (1997). *Fuzzy Logic and Neurofuzzy Applications in Business and Finance*, Prentice Hall, New York.
Wildt A. R., Mazis M. B. (1978). Determinants of Scale Response: Label versus Position, *Journal of Marketing Research*, 15, 261-267.
Zadeh L. A. (1965). Fuzzy sets, *Information and Control*, 8, 338-353.
Zimmermann H. J. (1996). *Fuzzy Set Theory and Its Applications*, third edition, Kluwer Academic, Boston.